Office de la propriété intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An Agency of Industry Canada

CA99/1008

Bureau canadien des brevets

Certification

Canadian Patent Office

Certification

La présente atteste que les documents ci-joints, dont la liste figure ci-dessous, sont des copies authentiques des documents déposés au Bureau des brevets.

This is to certify that the documents attached hereto and identified below are true copies of the documents on file in the Patent Office.

Specification and Drawings, as originally filed, with Application for Patent Serial No: **2„252,170**, on October 27, 1998, by **UNIVERSITE DE SHERBROOKE,** assignee of Bruno Bessette and Roch Lefebvre, for "A Method and Device for High Quality Coding of Wideband Speech and Audio Signals".

S. Grégoire

Agent certificateur/Certifying Officer

November 15, 1999

Date

Canada

(CIPO 68)

OPIC CIPO

A METHOD AND DEVICE FOR HIGH QUALITY CODING OF WIDEBAND SPEECH AND AUDIO SIGNALS

BACKGROUND OF THE INVENTION

1. Field of the invention:

The present invention relates to an efficient technique for digitally encoding a wideband sound signal, in particular but not exclusively a speech signal, in view of transmitting, or storing, and synthesizing this wideband sound signal.

2. Brief description of the prior art:

The demand for efficient digital wideband speech/audio encoding techniques with a good subjective quality/bit rate trade-off is increasing for numerous applications such as audio/video teleconferencing, multimedia, and wireless applications, as well as Internet and packet network applications. Until recently, telephone bandwidths filtered in the range 200-3400 Hz were mainly used in speech coding applications. However, there is an increasing demand for wideband speech applications in order to increase the intelligibility and naturalness of the speech signals. A bandwidth in the range 50-7000 Hz was found sufficient for delivering a face-to-face speech quality. For audio signals, this range gives an acceptable audio quality, but still lower than the CD quality which operates on the range 20-20000 Hz.

A speech encoder converts a speech signal into a digital bitstream which is transmitted over a communication channel (or stored in a storage medium). The speech signal is digitized (sampled and quantized with usually 16-bits per sample) and the speech encoder has the role of representing these digital samples with a smaller number of bits while maintaining a good subjective speech quality. The speech decoder or synthesizer operates on the transmitted or stored bit stream and converts it back to a sound signal.

One of the best prior art techniques capable of achieving a good quality/bit rate trade-off is the so-called Code Excited Linear Prediction (CELP) technique. According to this technique, the sampled speech signal is processed in successive blocks of *L* samples usually called *frames* where *L* is some predetermined number (corresponding to 10-30 ms of speech). In CELP, a linear prediction (LP) filter is computed and transmitted every frame. The *L*-sample frame is then divided into smaller blocks called *subframes* of size *N* samples, where *L=kN* and *k* is the number of subframes in a frame (*N* usually corresponds to 4-10 ms of speech). An excitation signal is determined in each subframe, which usually consists of two components: one from the past excitation (also called pitch contribution or adaptive codebook) and the other from an innovation codebook (also called fixed codebook). This excitation signal is transmitted and used at the decoder as the input of the LP synthesis filter in order to obtain the synthesized speech.

An innovation codebook in the CELP context, is an indexed set of *N*-sample-long sequences which will be referred to as *N*-dimensional codevectors. Each codebook sequence is indexed by an integer *k* ranging from 1 to *M* where *M* represents the size of the codebook often expressed as a number of bits b, where $M=2^b$.

To synthesize speech according to the CELP technique, each block of *N* samples is synthesized by filtering an appropriate codevector from a codebook through time varying filters modeling the spectral characteristics of the speech signal. At the encoder end, the synthetic output is computed for all, or a subset, of the codevectors from the codebook (codebook search). The retained codevector is the one producing the synthetic output closest to the original speech signal according to a perceptually weighted distortion measure. This perceptual weighting is performed using a so-called perceptual weighting filter, which is usually derived from the LP filter.

The CELP model has been very successful in encoding telephone band sound signals, and several CELP-based standards exist in a wide range of applications, especially in digital cellular applications. In the telephone band, the sound signal is band-limited to 200-3400 Hz and sampled at 8000 samples/sec. In wideband speech/audio applications, the sound signal is band-limited to 50-7000 Hz and sampled at 16000 samples/sec.

Some difficulties arise when applying the telephone-band optimized CELP model to wideband signals, and additional features need to be added to the model in order to obtain high quality wideband signals. Wideband signals exhibit a much wider dynamic range compared to telephone-band signals, which results in precision problems when a fixed-point implementation of the algorithm is required (which is essential in wireless applications). Further, the CELP model will often spend most of its encoding bits on the low-frequency region, which usually has higher energy contents, resulting in a low-pass output signal. To overcome this problem, the perceptual weighting filter has to be modified in order to suit wideband signals, and pre-emphasis techniques which boost the high frequency regions become important to reduce the dynamic range, yielding a simpler fixed-point implementation, and to ensure a better encoding of the higher frequency contents of the signal. Further, the pitch contents in the spectrum of voiced segments in wideband signals do not extend over the whole spectrum range, and the amount of voicing shows more variation compared to narrow-band signals. Thus, it is important to improve the closed-loop pitch analysis to better accommodate the variations in the voicing level.

At the decoder side, the CELP model uses post-filtering and post-processing techniques in order to improve the perceived synthesized signal. These techniques have to be changed to accomodate wideband signals. Further, in order to lower the bit rate below 16 kbit/s, an efficient method is to down-sample the wideband signals, which enables the encoder to operate on a bandwidth lower than 7000 Hz, thus achieving a reduction in the bit rate. At the decoder side, the decoder signal is upsampled and an efficient high frequency generation technique is needed to recover the full band signal, while maintaining a quality close to the original signal.

OBJECTS OF THE INVENTION

An object of the present invention is therefore to provide a method and device for efficiently encoding wideband (7000 Hz) sound signals using CELP-type encoding techniques, using additional features at both encoder and decoder in order to obtain high a quality reconstructed sound signal, which is also suitable for fixed point algorithmic implementation.

SUMMARY OF THE INVENTION

More specifically, in accordance with the present invention, there is provided a method for encoding wideband sound signals using LP-based, preferably CELP-type encoding techniques, whereby the following new features are adopted in order to obtain high subjective quality of the decoded wideband sound signal:

1. The overall perceptual weighting of the quantization error is obtained by a combination of a preemphasis filter and a modified weighting filter.

In CELP-type coders, the optimum pitch and innovation parameters are searched by minimizing the mean squared error between the input speech and synthesized speech in a perceptually weighted domain. This is equivalent to minimizing the error between the weighted input speech and weighted synthesis speech, where the weighting is performed using a filter having a transfer function $W(z)$ of the form:

$$W(z)=A(z/\gamma_1)/A(z/\gamma_2) \quad \text{where} \quad 0<\gamma_2<\gamma_1\leq 1$$

In analysis-by-synthesis (AbS) coders, analysis show that the quantization error is weighted by the inverse of the weighting filter, $W^{-1}(z)$, which exhibits some of the formant structure in the input signal. Thus, the masking property of the human ear is exploited by shaping the error, so that it has more energy in the formant regions, where it will be masked by the strong signal energy present in those regions. The amount of weighting is controlled by the factors γ_1 and γ_2.

This filter works well with telephone band signals. However, it was found that this filter is not suitable for efficient perceptual weighting when it was applied to wideband signals. It was found that this filter has inherent limitations in modeling the formant structure and the required spectral tilt concurrently. The spectral tilt is more pronounced in wideband signals due to the wide dynamic range between low and high frequencies. It was suggested to add a tilt filter into filter *W*(z) in order to control the tilt and formant weighting separately.

A novel solution to this problem, forming part of the present invention, is to introduce a preemphasis filter at the input, compute the LP filter *A*(z) based on the preemphasized speech, and use a modified filter *W*(z) by fixing its denominator.

The preemphasis filter reduces the dynamic range of the input signal, which renders it more suitable for fixed-point implementation, and improves the encoding of the high frequency contents of the spectrum. The preemphasis is obtained by a fixed FIR filter having a transfer function $P(z)$ in the form:

$$P(z)=1-\mu z^{-1}$$

where μ is a preemphasis factor with a value between 0 and 1. A higher order filter can also be used. Linear prediction (LP) analysis is performed on the preemphasized input signal to obtain the LP filter $A(z)$. A new weighting filter is used, which has a transfer function of the form:

$$W(z)=A(z/\gamma_1)/(1-\gamma_2 z^{-1}) \qquad \text{where } 0<\gamma_2<\gamma_1\leq 1$$

Note that because $A(z)$ is computed based on preemphasized speech, the tilt of the filter $1/A(z/\gamma_1)$ is less pronounced compared to the case when $A(z)$ is computed based on the original speech. Since deemphasis using the filter $P^{-1}(z)=1/(1-\mu z^{-1})$ is performed at the receiver end, the quantization error spectrum is shaped by the filter $W^{-1}(z)P^{-1}(z)$. When μ is set equal to γ_2, which is typically the case, the spectrum of the quantization error is shaped by the filter $1/A(z/\gamma_1)$, with $A(z)$ computed based on the preemphasized speech. Subjective listening showed that this structure of achieving the error shaping by a combination of preemphasis and modified weighting filtering is very efficient for encoding wideband signals, in addition to the advantages of ease of fixed-point algorithmic implementation.

2. The closed-loop pitch analysis is improved to better accommodate wideband signals.

The pitch harmonics in AbS coders are usually modeled using a pitch delay *T* and an associated gain *b*. The excitation signal *u*(*n*) is derived by adding the past excitation at delay *T* scaled by a gain *b* to an innovation component from a fixed codebook scaled by a gain *g*. That is

$$u(n)=bv_T(n)+gc_k(n)$$

where $v_T(n)$ is the past excitation at delay *T* samples. For an improved performance, a fractional delay is usually used. In this case, the past excitation is oversampled to achieve the required higher resolution. In most cases, the pitch predictor can be represented by a filter having a transfer function of the form $1/(1-bz^{-T})$, whose spectrum has a harmonic structure over the entire frequency range, with a harmonic frequency related to 1/*T*. In case of wideband signals, this structure is not very efficient since the harmonic frequencies don't cover the entire extended spectrum. The harmonic structure exists only up to a certain frequency, depending on the speech segment. A new method which achieves efficient modeling of the harmonic structure of the speech spectrum uses several forms of low pass filters applied to the past excitation and the one yielding higher prediction gain is selected. When subsample pitch resolution is used, the low pass filters can be incorporated into the interpolation filters used to obtain the higher pitch resolution.

3. At the decoder, the innovative contribution to the excitation is enhanced by filtering it through a preemphasis filter whose coefficients are derived from the level of voicing in speech segement in the subframe.

Enhancing the periodicity of the excitation signal improves the quality in case of voiced segments. This was done in the past by filtering the innovation from the fixed codebook through a filter having a transfer function of the form $1/(1-\varepsilon z^{-T})$ where ε is a factor below 0.5 which controls the amount of introduced periodicity. This approach is less efficient in case of wideband signals since it introduces the periodicity over the entire spectrum. A new alternative approach is disclosed whereby the periodicity enhancement is achieved by filtering the innovative signal from the fixed codebook by a filter which emphasizes the high frequencies and reduces the low-frequency contents of the innovation, and whose coefficients are related to the level of periodicity in the signal. In this approach, the innovative contribution is reduced mainly at low frequencies, which enhances the periodicity of the excitation at low frequencies more than high frequencies.

4. A new high-frequency generation procedure is introduced in order to recover the high frequency content of the signal, in case the input signal has been down-sampled.

In order to improve the coding efficiency and reduce the algorithmic complexity of the wideband coding algorithm, the input wideband signal is down-sampled from 16 kHz to around 12.8 kHz. This reduces the number of samples in a frame which reduces the processing time, and reduces the signal bandwidth which enables the reduction in bit rate down to 12 kbit/s while keeping very high quality decoded sound signal. At the decoder, the high frequency contents of the signal needs to be reintroduced to remove the low pass filtering effect from the decoded signal and retrieve the natural sounding quality of wideband signals. A new approach consists of generating the high frequency contents by filling the upper part of the spectrum with a white noise properly scaled in the excitation domain, then converted to the speech domain, preferably but not necessarily by shaping it with the same LP filter used for synthesizing the down-sampled signal.

The objects, advantages and other features of the present invention will become more apparent upon reading of the following non restrictive description of a preferred embodiment thereof, given by way of example only with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

In the appended drawings:

Figure 1 is a schematic block diagram of a preferred embodiment of a wideband encoding device embodying the present invention;

Figure 2 is a schematic block diagram of a preferred embodiment of a wideband decoding device embodying the present invention, and comprising a method for high frequency generation; and

Figure 3 is a schematic block diagram of a closed-loop pitch analysis device suitable for wideband signals.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The novel techniques disclosed in the present specification may apply to different LP (Linear Prediction)-based coding systems. However, a CELP-type coding system is used in the preferred embodiment for presenting a non limitative illustration of the techniques disclosed herein.

Figure 1 shows a general block diagram of a CELP-type speech encoding device modified to better accommodate wideband signals.

The sampled input speech is divided into L-sample blocks called "frames". In each frame, different parameters representing the speech signal in the frame are computed, encoded, and transmitted. LP parameters representing the LP synthesis filter are usually computed once every frame. The frame is further divided into smaller blocks of length N, in which excitation parameters (pitch and innovation) are determined. In the CELP literature, these blocks of length N are called "subframes" and the N-sample signals in a subframe are referred to as N-dimensional vectors. In this preferred embodiment, the length N corresponds to 5 ms while the length L corresponds to 20 ms, which means that a frame contains four subframes (N=80 at the sampling rate of 16 kHz and 64 after down-sampling to 12.8 kHz). Various N-dimensional vectors occur in the encoding procedure. A list of the vectors which appear in Figures 1 and 2 as well as a list of transmitted parameters are given herein below:

List of the main *N*-dimensional vectors

$\boldsymbol{s}$ Input speech vector (after down-sampling, pre-processing, and preemphasis);

$\mathbf{s}_w$ Weighted speech vector;

$\mathbf{s}_0$ Zero-input response of weighted synthesis filter;

$\mathbf{x}$ Target vector for pitch search;

$\mathbf{h}$ Impulse response of the combination of synthesis and weighting filters;

$\mathbf{v}_T$ Adaptive codebook vector at delay T;

$\mathbf{y}_T$ Filtered adaptive codebook vector ($\mathbf{v}_T$ convolved with $\mathbf{h}$);

$\mathbf{x}'$ Target vector for pitch search;

$\mathbf{c}_k$ Innovation codevector at index k (k-th entry from the innovation codebook);

$\mathbf{c}_f$ Enhanced scaled innovation codevector;

$\mathbf{u}$ Excitation signal (scaled innovation and pitch codevectors);

$\mathbf{u}'$ Enhanced excitation;

$\mathbf{s}'$ Synthesis signal before deemphasis; and

$\mathbf{s}_h$ Synthesis signal after deemphasis and postprocessing.

List of transmitted parameters

STP Short term prediction parameters (defining $A(z)$);

T Pitch lag (or adaptive codebook index);

b Pitch gain (or adaptive codebook gain);

j	Index of the low-pass filter used on the pitch codevector;
k	Codevector index (innovation codebook entry); and
g	Innovation codebook gain.

In this preferred embodiment, the STP parameters are transmitted once per frame and the rest of the parameters are transmitted four times per frame (every subframe).

ENCODING PRINCIPLE

The sampled speech signal is encoded on a block by block basis by the encoding device of Figure 1 which is broken down into eleven modules numbered from 101 to 111.

The input speech is processed into the above mentioned *L*-sample blocks called frames.

Referring to Figure 1, the input speech signal is down-sampled in a down-sampling module 101. In this preferred embodiment, the signal is down-sampled from 16 kHz down to 12.8 kHz, using techniques well known in the art. Down-sampling increases the coding efficiency, since a smaller bandwidth is encoded. This also reduces the algorithmic complexity since the number of samples in a frame is decreased. The use of down-sampling becomes significant as the bit rate is reduced below 16 kbit/s, although down-sampling is not essential above 16 kbit/s.

After down-sampling, the 320-sample frame of 20 ms is reduced to 256-sample frame (down-sampling ratio of 4/5).

The input frame is then passed into the optional pre-processing block 102, which consists of a high pass filter with a 50 Hz cut-off frequency. High-pass filter 102 removes the unwanted sound components below 50 Hz.

The down-sampled pre-processed signal is denoted by $s_p(n)$, n=0,...,L-1, where L is the length of the frame (256 at 12.8 kHz sampling). In preemphasis 103, the signal $s_p(n)$ is preemphasized using a filter having the following transfer function:

$$P(z)=1-\mu z^{-1}$$

where μ is a preemphasis factor with a value between 0 and 1 (a typical value is μ=0.7). A higher order filter can also be used.

Note that the high-pass filter 102 and preemphasis filter 103 can be interchanged to obtain more efficient fixed-point implementations.

The function of the preemphasis filter 103 is to reduce the dynamic range of the input speech signal, which renders it more suitable for fixed-point implementation. Without preemphasis, it is difficult to implement LP analysis in fixed-point using single-precision arithmetic.

Preemphasis also plays an important role in achieving a proper overall perceptual weighting of the quantization error, which contributes to an improved sound quality. This will be explained later in more details.

The output of the preemphasis filter 103 is denoted *s*(*n*). This signal is used for performing LP analysis, a technique well known in the art. The autocorrelation approach is used, where the signal is first windowed using a Hamming window (usually in the order of 30-40 ms). The autocorrelations are computed from the windowed signal, and Levinson-Durbin recursion is used to compute the LP parameters, a_i, where i=1,...,p, and where p is the LP order, which is typically 16 in wideband coding. The parameters a_i are the coefficients of the transfer function of the LP filter:

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$$

LP analysis is performed in module 104, which also performs the quantization and interpolation of the LP parameters. The LP coefficients are transformed into another equivalent domain more suitable for quantization and interpolation purposes. The line spectral pair (LSP) and immitance spectral pair (ISP) domains are two domains in which quantization and interpolation can be efficiently performed. The 16 LP parameters can be quantized in the order of 30 to 50 bits using split or multi-stage quantization, or a combination thereof. The purpose of the interpolation is to enable updating the LP parameters every subframe while transmitting them once every frame, which improves the coder performance without increasing the bit rate.

The following paragraphs will describe the rest of the coding operations performed on a subframe basis. In the following description, the filter $A(z)$ denotes the unquantized interpolated LP filter in the subframe, and the filter $\hat{A}(z)$ denotes the quantized interpolated LP filter in the subframe.

Perceptual Weighting:

In analysis-by-synthesis coders, the optimum pitch and innovation parameters are searched by minimizing the mean squared error between the input speech and synthesized speech in a perceptually weighted domain. This is equivalent to minimizing the error between the weighted input speech and weighted synthesis speech.

The weighted signal $s_w(n)$ is computed in a weighted signal generator 105. Traditionally, the weighted signal $s_w(n)$ is computed by a weighting filter having a transfer function $W(z)$ in the form

$$W(z)=A(z/\gamma_1)/A(z/\gamma_2) \quad \text{where} \quad 0<\gamma_2<\gamma_1\leq 1$$

In analysis-by-synthesis (AbS) coders, analysis shows that the quantization error is weighted by a transfer function, $W^{-1}(z)$, which is the inverse of the transfer function of the filter 105. Transfer function $W^{-1}(z)$ exhibits some of the formant structure in the input signal. Thus, the masking property of the human ear is exploited by shaping the error, so that it has more energy in the formant regions, where it will be masked by the strong signal energy present in those regions. The amount of weighting is controlled by the factors γ_1 and γ_2.

The above traditional weighting filter works well with telephone band signals. However, it was found that this weighting filter is not suitable for efficient perceptual weighting when it was applied to wideband signals. It was found that this filter has inherent limitations in modeling the formant structure and the required spectral tilt concurrently. The spectral tilt is more pronounced in wideband signals due to the wide dynamic range between low and high frequencies. The prior art has suggested to add a tilt filter into *W*(*z*) in order to control the tilt and formant weighting separately.

A novel solution to this problem, which is part of the present invention, is to introduce the preemphasis filter 103 at the input, compute the LP filter *A*(*z*) based on the preemphasized speech *s*(*n*), and use a modified filter *W*(*z*) by fixing its denominator.

LP analysis is performed in module 104 on the preemphasized signal *s*(*n*) to obtain the LP filter *A*(*z*). A new perceptual weighting filter 105 with fixed denominator

$$W(z)=A(z/\gamma_1)/(1-\gamma_2 z^{-1}) \qquad \text{where } 0<\gamma_2<\gamma_1\leq 1$$

is used (a higher order can be used at the denominator). This form decouples the formant weighting from the tilt.

Note that because *A*(*z*) is computed based on the preemphasized speech signal *s*(*n*), the tilt of the filter $1/A(z/\gamma_1)$ is less pronounced compared to the case when *A*(*z*) is computed based on the original speech. Since deemphasis is made at the receiver end using a filter having a transfer function $P^{-1}(z)=1/(1-\mu z^{-1})$, the quantization error spectrum is shaped by a filter having a transfer function $W^{-1}(z)P^{-1}(z)$. When μ is set equal to γ_2, which is typically the case, the spectrum of the quantization error is shaped by a filter whose transfer function is $1/A(z/\gamma_1)$, with $A(z)$ computed based on the preemphasized speech. Subjective listening showed that this structure of achieving the error shaping by a combination of preemphasis and modified weighting filtering is very effcicient for encoding wideband signals, in addition to the advantages of ease of fixed-point algorithmic implementation.

Pitch Analysis:

In order to simplify the pitch analysis, an open-loop pitch lag is first estimated in the open-loop pitch search module 106 using the weighted speech signal $s_w(n)$. Then the closed-loop pitch analysis which is performed in closed-loop pitch search module 107 on a subframe basis is restricted around the open-loop pitch lag which significantly reduces the search complexity of the LTP parameters *T* and *b* (pitch lag and pitch gain). Open-loop pitch analysis is usually performed once every 10 ms (two subframes) using techniques well known in the art.

The target signal for LTP (Long Term Prediction) analysis, **x**, is first computed. This is usually done by subtracting the zero-input response of a weighted synthesis filter $W(z)/\hat{A}(z)$ (calculated by a zero-input response generator 108) from the weighted speech signal s_w *(n)*. More specifically, the target vector ***x*** is calculated using the following relation:

$$\mathbf{x}=\mathbf{s}_w-\mathbf{s}_0$$

where ***x*** is the *N*-dimensional target vector, $\mathbf{s}_w$ is the weighted signal vector in the subframe, and $\mathbf{s}_0$ is the zero-input response of the filter *W(z)/Â(z)* which is the output of the combined filter *W(z)/Â(z)* due to its initial states. $\mathbf{s}_0$ is computed in the zero-input response generator 108.

Just a word to mention that alternative, but mathematically equivalent approaches can be used to compute the target vector.

A *N*-dimensional impulse response vector **h** of the weighted synthesis filter *W(z)/Â(z)* is computed in the impulse response generator 109.

The closed-loop pitch or adaptive codebook parameters are computed in the closed-loop pitch search module 107, which uses the target vector **x** and the impulse response vector **h** as inputs. Traditionally, the pitch prediction was represented by a pitch filter having the following transfer function:

$$1/(1-bz^{-T})$$

where *b* is the pitch gain and *T* is the pitch delay or lag. In this case, the pitch contribution to the excitation signal *u*(*n*) is given by $bu(n-T)$, where the total excitation is given by

$$u(n)=bu(n-T)+gc_k(n)$$

with g being the innovative codebook gain and $c_k(n)$ the innovation codevector at index k.

This representation has limitations if the delay T is shorter than the subframe length N. In another view point, the pitch contribution can be seen as an adaptive codebook containing the past excitation signal. Generally, each vector in the adaptive codebook is a shift-by-one version of the previous vector (discarding one sample and adding a new sample). For delays $T>N$, the adaptive codebook is equivalent to the filter structure, and a codevector $v_T(n)$ is given by

$$v_T(n)=u(n-T), \qquad n=0,\ldots,N\text{-}1.$$

For delays shorter than T, a codevector is built by repeating the available samples from the past excitation until the codevector is completed (this is not equivalent to the filter structure).

In recent coders, a higher pitch resolution is used which significantly improves the quality of voiced sound segments. This is achieved by oversampling the past excitation signal using polyphase interpolation filters. In this case, the codevector $v_T(n)$ may correspond to an interpolated version of the past excitation, with T being a non-integer delay (e.g. 50.25).

The pitch search consists of finding the best delay T and gain b that minimize the mean squared weighted error between the target vector $\boldsymbol{x}$ and the scaled filtered past excitation

$$E = \|\mathbf{x} - b\mathbf{y}_T\|^2$$

where $\mathbf{y}_T$ is the filtered adaptive codevector at delay T:

$$y_T(n) = v_T(n) * h(n) = \sum_{i=0}^{n} v_T(i)h(n-i), \qquad n=0,\ldots,N-1.$$

It can be shown that the error E is minimized by maximizing the criterion

$$C = \frac{\mathbf{x}^t \mathbf{y}_T}{\sqrt{\mathbf{y}_T^t \mathbf{y}_T}}$$

where t denotes vector transpose.

In the preferred embodiment of the present invention, a 1/3 subsample pitch resolution is used, and the pitch search is composed of three stages.

In the first stage, an open-loop delay is estimated in open-loop pitch search module 106. In the second stage, the search criterion C is seached in the closed-loop pitch search module 107 for integer delays around the estimated open-loop delay (usually ±5), which significantly simplifies the search procedure. A simple procedure is used for updating the filtered codevector $\mathbf{y}_T$ without the need to compute the convolution for every delay. Once an optimum integer delay is found, the fractions around the integer delay are tested in the third stage of the search (module 107).

When the pitch predictor is represented by a filter of the form $1/(1-bz^{-T})$, which is a valid assumption for delays $T>N$, the spectrum of the pitch filter exhibits a harmonic structure over the entire frequency range, with a harmonic frequency related to $1/T$. In case of wideband signals, this structure is not very efficient since the harmonic structure in wideband signals does not cover the entire extended spectrum. The harmonic structure exists only up to a certain frequency, depending on the speech segment. Thus, in order to achieve efficient representation of the pitch contribution in voiced segments of wideband speech, the pitch predictor need to have the flexibility of varying the amount of periodicity over the wideband spectrum.

A new method which achieves efficient modeling of the harmonic structure of the speech spectrum is disclosed in the present specification, whereby several forms of low pass filters are applied to the past excitation and the one with higher prediction gain is selected.

When subsample pitch resolution is used, the low pass filters can be incorporated into the interpolation filters used to obtain the higher pitch resolution. In this case, the third stage of the pitch search, in which the fractions around the chosen integer delay are tested, is repeated for the several interpolation filters having different low-pass characteristics and the fraction and filter index which maximize the search criterion C are selected.

A simpler approach, is to complete the search in the three stages described above, to determine the optimum fractional delay using only one interpolation filter with certain frequency response, and select the optimum low-pass filter shape at the end by applying the different predetermined low-pass filters to the chosen adaptive codevector $\mathbf{v}_T$ and select the low-pass filter which minimizes the pitch prediction error.

Figure 3 shows a schematic block diagram of a preferred embodiment of the proposed approach.

In module 303, the past excitation codevector is memorized. Module 301 is responsive to the target vector $\boldsymbol{x}$ and to the past excitation codevector from memory module 303 to conduct a pitch codebook search minimizing the above-defined search criterion C. From the result of the search conducted in module 301, module 302 generates the optimum codevector $\boldsymbol{v}_T$.

Suppose that K filter characteristics are used (they could be low-pass or band-pass). Once the optimum codevector $\mathbf{v}_T$ is determined, K filtered versions of $\mathbf{v}_T$ are computed using the K different frequency shaping filters such as 305$^{(j)}$, where j=1, ... , K. These filtered versions are denoted $\mathbf{v}_f^{(j)}$, j=1,...,K. The different vectors $\mathbf{v}_f^{(j)}$ are convolved in modules 304$^{(j)}$, where j=1, ... , K, with the impulse response **h** to obtain the vectors $\mathbf{y}^{(j)}$, j=1,...,K. The selected frequency shaping filter 305$^{(j)}$ is the one which minimizes the mean squared pitch prediction error

$$e^{(j)} = \left\| \mathbf{x} - b^{(j)} \mathbf{y}^{(j)} \right\|^2, \quad j=1,\ldots,K$$

To calculate the mean squared pitch prediction error for each value of $\boldsymbol{y}^{(j)}$, the value $y^{(j)}$ is multiplied by the gain b by means of an amplifier $307^{(j)}$ and the value $b^{(j)}y^{(j)}$ is subtracted from the target vector x by means of subtractors $308^{(j)}$.

The gain $b^{(j)}$ associated with the frequency shaping filter at index j, is given by

$$b^{(j)} = \mathbf{x}^t\mathbf{y}^{(j)} / \left\|\mathbf{y}^{(j)}\right\|^2.$$

In the same manner, optimum codevector $\mathbf{v}_T$ is convolved with the impulse response **h** to obtain the vectors y. To calculate the mean squared pitch prediction error for y, the value y is multiplied by the gain b by means of an amplifier $307^{(j)}$ and the value by is subtracted from the target vector x by means of subtractors 308. The gain b is given by

$$b = x^t\, y / \|y\|^2$$

In module 309, the parameters b, T, and j are chosen based on v_T or $v_f^{(j)}$ which minimizes the mean squared pitch prediction error e.

The pitch codebook index T is encoded and transmitted. The pitch gain b is quantized and transmitted. With this new approach, extra information is needed to encode the index j of the selected frequency shaping filter. If two filters are used, then one bit is needed to represent this information.

Innovative codebook search:

Once the pitch, or LTP (Long Term Prediction) parameters *b*, *T*, and *j* are determined, we proceed by searching for the optimum innovative excitation by means of module 110 of Figure 1. First, the target vector **x** is updated by subtracting the LTP contribution:

$$\mathbf{x}' = \mathbf{x} - b\mathbf{y}_T$$

where *b* is the pitch gain and $\mathbf{y}_T$ is the filtered adaptive codebook vector (the past excitation at delay *T* filtered with the selected low pass filter and convolved with the inpulse response **h** as described with reference to Figure 3).

The search procedure in CELP is performed by finding the optimum excitation codevector $\mathbf{c}_k$ and gain *g* which minimize the mean-squared error between the target vector and the scaled filtered codevector

$$E = \left\| \mathbf{x}' - g\mathbf{H}\mathbf{c}_k \right\|^2$$

where **H** is a lower triangular convolution matrix derived from the impulse response vector **h**.

In the preferred embodiment of the present invention, the innovative codebook search is performed in module 110 by means of an algebraic codebook as described in US patent numbers 5,444,816 (Adoul et al.) issued on August 22, 1995; 5,699,482 granted to Adoul et al., on December

17, 1997; 5,754,976 granted to Adoul et al., on May 19, 1998; and 5,701,392 (Adoul et al.) dated December 23, 1997.

Once the optimum codevector and its gain are chosen by module 110, the codebook index *k* and gain *g* are encoded and transmitted.

Referring to Figure 1, the parameters *b, T, j, Â(z), k* and *g* are multiplexed through a multiplexer 112 before being encoded and tranmitted

Memory update:

In module 111 (Figure 1), the states of the weighted synthesis filter are updated by filtering the excitation signal $\mathbf{u}=g\mathbf{c}_k+b\mathbf{v}_T$ through the weighted synthesis filter. At the end of this filtering, the states of the filter are memorized and used in the next subframe as initial states for computing the zero-input response in generator module 108.

Similar to the target vector, other alternative, but mathematically equivalent, approaches can be used to update the filter states.

DECODING PRINCIPLE

The speech decoding device of Figure 2 illustrates the various steps carried out between the digital input 222 (input to the demultiplexer 217) and the output sampled speech 223 (output of the adder 221).

The demultiplexer 217 extracts the synthesis model parameters from the binary information received from a digital input channel. From each received binary frame, the extracted parameters are:

- the short-term prediction parameters STP (once per frame);

- the long-term prediction (LTP) parameters *T*, *b*, and *j* (for each subframe); and

- the innovation codebook index *k* and gain *g* (for each subframe).

The current speech signal is synthesized based on these parameters as will be explained hereinbelow.

The innovative excitation generator 218 is responsive to the index *k* to produce the innovation codevector $\mathbf{c}_k$, which is scaled by the decoded gain factor *g* through an amplifier 224. In the preferred embodiment, an algebraic codebook as described in the above mentioned US patent numbers 5,444,816; 5,699,482; 5,754,976; and 5,701,392 is used to represent the innovative excitation.

The generated scaled codevector at the output of the amplifier 224 is processed through a frequency-dependent pitch enhancer 205.

Enhancing the periodicity of the excitation signal improves the quality in case of voiced segments. This was done in the past by filtering the innovation from the fixed codebook through a filter in the form $1/(1-\varepsilon z^{-T})$ where ε is a factor below 0.5 which controls the amount of introduced periodicity. This approach is less efficient in case of wideband signals since it introduces the periodicity over the entire spectrum. A new alternative approach, which is part of the present invention, is disclosed whereby the periodicity enhancement is achieved by filtering the innovative signal from the fixed codebook by a filter *F*(*z*) whose frequency response emphasizes the higher frequencies more than lower frequencies. The coefficients of *F*(*z*) are related to the amount of periodicity in the signal. An efficient way to derive the filter coefficients is to relate them to the amount of pitch contribution to the total excitation. This results in a frequency response depending on the subframe periodicity, where higher frequencies are more strongly emphasized (stronger overall slope) for higher pitch gains. This filter has the effect of lowering the energy of the innovative excitation at low frequencies when the signal is more periodic, which enhances the periodicity of the excitation at lower frequencies more than higher frequencies. Suggested forms of this filter are

(1) $F(z)=1-\sigma z^{-1}$ or (2) $F(z)=-\alpha z+1-\alpha z^{-1}$

where σ or α are factors derived from the level of periodicity of the signal. The second 3-tape form of *F*(*z*) is used in this preferred embodiment. The factor α is computed in the voicing factor generator 204 as follows:

The ratio of pitch contribution to the total excitation is first computed by

$$R_p=\frac{b^2\mathbf{v}_T^t\mathbf{v}_T}{\mathbf{u}^t\mathbf{u}}=\frac{b^2\sum_{n=0}^{N-1}v_T^2(n)}{\sum_{n=0}^{N-1}u^2(n)}$$

where $\mathbf{v}_T$ is the pitch codebook vector, b is the pitch gain, and $\mathbf{u}$ is the excitation vector given at the output of the adder 219 by

$$\mathbf{u} = b\mathbf{v}_T + g\mathbf{c}_k$$

Just a word to mention that the term $b\boldsymbol{v}_T$ is produced by the pitch codebook 201 in response to the pitch lag T and the past value of $\boldsymbol{u}$ stored in memory 203. The adaptive codevector from the pitch codebook 201 is then processed through a low-pass filter whose cut-off frequency is adjusted by means of the index j from the demultiplexer 217. The resulting codevector $\boldsymbol{v}_T$ is then multiplied by the gain g from the demultiplexer 217 through an amplifier 226 to obtain the signal $b\boldsymbol{v}_T$.

The factor α is given by

$\alpha = qR_p$ bounded by $\alpha < q$

where q is a factor which controls the amount of enhancement (q is set to 0.25 in this preferred embodiment).

The enhanced signal $\mathbf{c}_f$ is computed by filtering the scaled innovative vector $g\mathbf{c}_k$ through the enhancing filter $F(z)$.

The enhanced excitation signal $\mathbf{u}'$ is computed by the adder 220 as

$$\mathbf{u}' = b\mathbf{v}_T + \mathbf{c}_f$$

Note that this process is not performed at the encoder. Thus, it is essential to update the content of the adaptive codebook using the excitation without enhancement to keep synchronism between the encoder and decoder. Therefore, the excitation signal **u** is used to update the memory of the adaptive codebook and the enhaced excitation signal **u**' is used at the input of the LP synthesis filter 206.

The synthesized signal **s**' is computed by filtering the enhanced excitation signal **u**' through the LP synthesis filter 206 which has the form $1/\hat{A}(z)$, where $\hat{A}(z)$ is the interpolated LP filter in the current subframe. As can be seen in Figure 2, the LP coefficients 225 from the demultiplexer 217 are supplied to the LP filter 206 to adjust the parameters of the LP filter 206 accordingly. The deemphasis filter 207 is the inverse of the preemphasis filter 103 of Figure 1. The transfer function of the preemphasis filter 108 is given by

$$D(z)=1/(1-\mu z^{-1})$$

The vector **s**' is filtered through the deemphasis filter *D*(*z*) (module 207) to obtain the vector $\mathbf{s}_d$, which is passed through the postprocessing module 208 comprising a high-pass filter to remove the unwanted frequencies below 50 Hz.

The over-sampling module 209 conducts the inverse process of the down-sampling module 101 of Figure 1. In this preferred embodiment, oversampling converts from the 12.8 kHz sampling rate to the original 16 kHz sampling rate, using techniques well known in the art. The oversampled synthesis signal is denoted $\hat{s}$.

The synthesis signal does not contain the higher frequency components which were lost by the downsampling process (module 101 of Figure 1) at the encoder. This gives a low-pass perception of the synthesis speech. To restore the full band of the original signal, a high frequency generation procedure is disclosed. This procedure is performed in modules 212 through 216 of Figure 2.

In this new approach, the high frequency contents are generated by filling the upper part of the spectrum with a white noise properly scaled in the excitation domain, then converted to the speech domain, preferably by shaping it with the same LP filter used for synthesizing the down-sampled signal.

The high frequency generation procedure, which is part of the present invention, is detailed hereinbelow.

The random noise generator 213 generates a white noise sequence **w'** with a flat spectrum over the entire frequency bandwidth, using techniques well known in the art. The generated sequence is of length *N'* which is the subframe length in the original domain. Note that *N* is the subframe length in the down-sampled domain. In this preferred embodiment, *N*=64 and *N'*=80 which correspond to 5 ms.

The white noise sequence is properly scaled in the gain adjusting module 214. Gain adjustment comprises the following steps. First, the energy of the generated noise sequence is set equal to the energy of the enhanced excitation signal **u'** computed by an energy computing module 210, and the resulting scaled noise sequence **w** is given by

$$w(n) = w'(n)\sqrt{\frac{\sum_{n=0}^{N-1} u^2(n)}{\sum_{n=0}^{N-1} w'^2(n)}}, \qquad n=0,\ldots,N'-1$$

The second step in the gain scaling is to take into account the voicing of the synthesized signal at the output of generator 204 so as to reduce the energy of the generated noise proportional to the voicing. In this preferred embodiment, this is implemented by measuring the tilt of the synthesis signal through a spectral tilt calculator 212 and reducing the energy accordingly. When the tilt is very strong, which corresponds to voiced segments, the noise energy is further reduced. The tilt factor is computed in module 212 as the first correlation coefficient of the synthesis signal $\mathbf{s}_h$ and it is given by

$$tilt = \frac{\sum_{n=1}^{N-1} s_h(n)s_h(n-1)}{\sum_{n=0}^{N-1} s_h^2(n)}, \qquad \text{bounded by} \quad tilt \geq 0 \text{ and } tilt \geq r_v.$$

r_v is given by

$r_v = (E_v - E_c)/(E_v + E_c)$ where E_v is the energy of the scaled pitch codevector and E_c is the energy of the scaled innovative codevector. r_v is mostly less than *tilt* but this bound was introduced as a precaution against high frequency tones where the tilt value is high and the value of r_v is small. So this bound reduces the noise energy for such tonal signals.

The tilt value is 0 in case of flat spectrum and 1 in case of strongly voiced signals. The scaling factor derived from the tilt is given by

$$g_t = 10^{-0.6 tilt}$$

When the tilt is close to zero, the scaling factor is close to 1, which does not result in energy reduction. When the tilt value is 1, the scaling factor results in a reduction of 12 dB in the energy of the generated noise.

Once the noise is properly scaled, it is brought into the speech domain using the spectral shaper 215. In the preferred embodiment, this is achieved by filtering the noise through a bandwidth expanded version of the same LP synthesis filter used in the down-sampled domain (1/Â(z/0.8)).

The filtered scaled noise sequence is then band-pass filtered to the required frequency range to be restored using the band-pass filter 216. In the preferred embodiment, the band-pass filter 216 restricts the noise sequence to the frequency range 5.6-7.2 kHz. The resulting band-pass noise sequence **z** is added to the oversampled synthesized speech signal $\hat{s}$ to obtain the final reconstructed sound signal s_{out} on the output 223.

Although the present invention has been described hereinabove by way of a preferred embodiment thereof, this embodiment can be modified at will, within the scope of the appended claims, without departing from the spirit and nature of the subject invention.

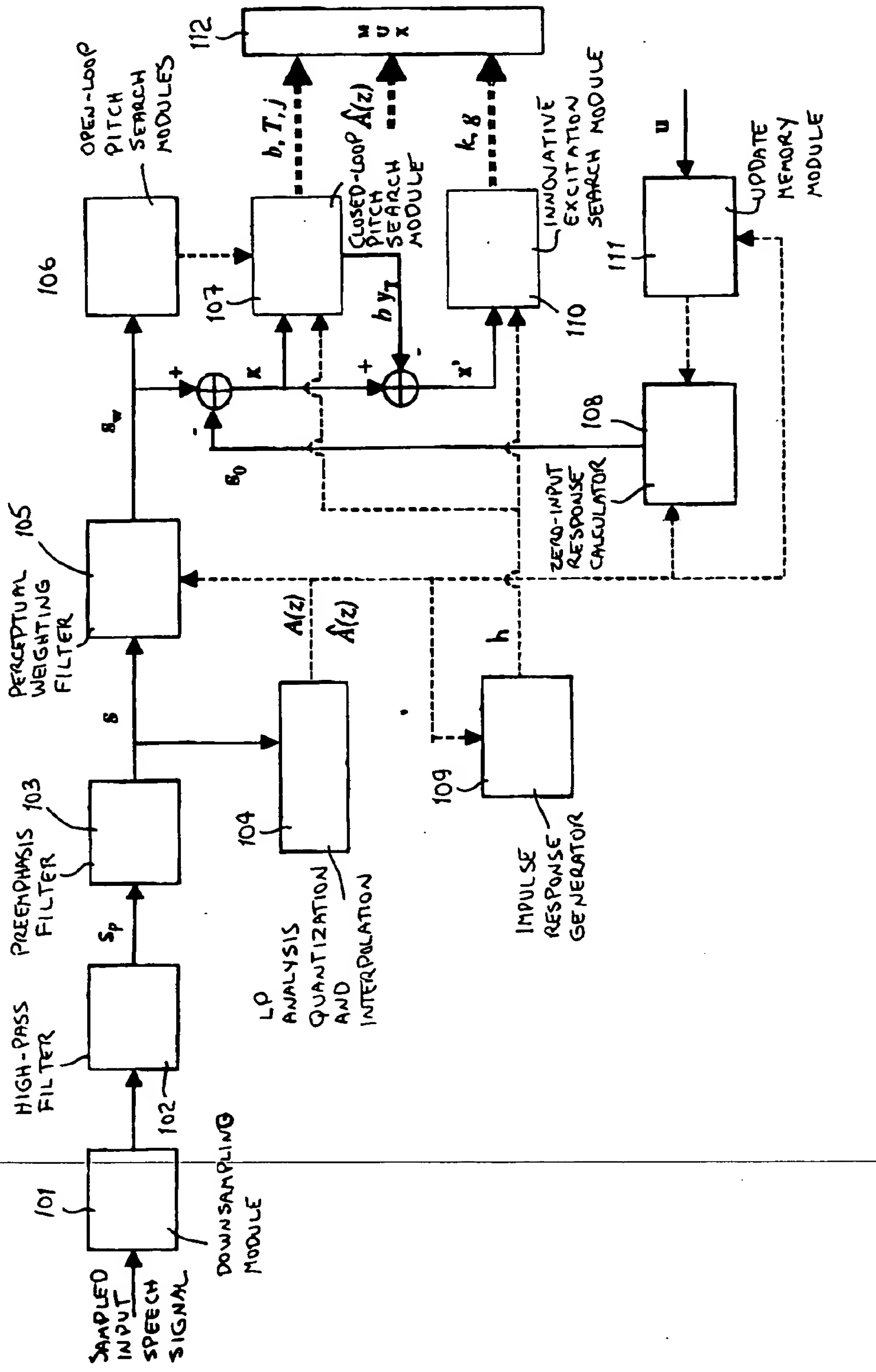
SAMPLED INPUT SPEECH SIGNAL
101
DOWNSAMPLING MODULE
102
HIGH-PASS FILTER
s_p
103
PREEMPHASIS FILTER
s
104
LP ANALYSIS QUANTIZATION AND INTERPOLATION
$A(z)$
$\hat{A}(z)$
105
PERCEPTUAL WEIGHTING FILTER
s_w
106
OPEN-LOOP PITCH SEARCH MODULES
107
CLOSED-LOOP PITCH SEARCH MODULE
b, T, j
$\hat{A}(z)$
$b y_T$
x
x'
s_0
+
-
+
-
108
ZERO-INPUT RESPONSE CALCULATOR
109
IMPULSE RESPONSE GENERATOR
h
110
INNOVATIVE EXCITATION SEARCH MODULE
k, g
111
UPDATE MEMORY MODULE
u
112
MUX

Figure 1

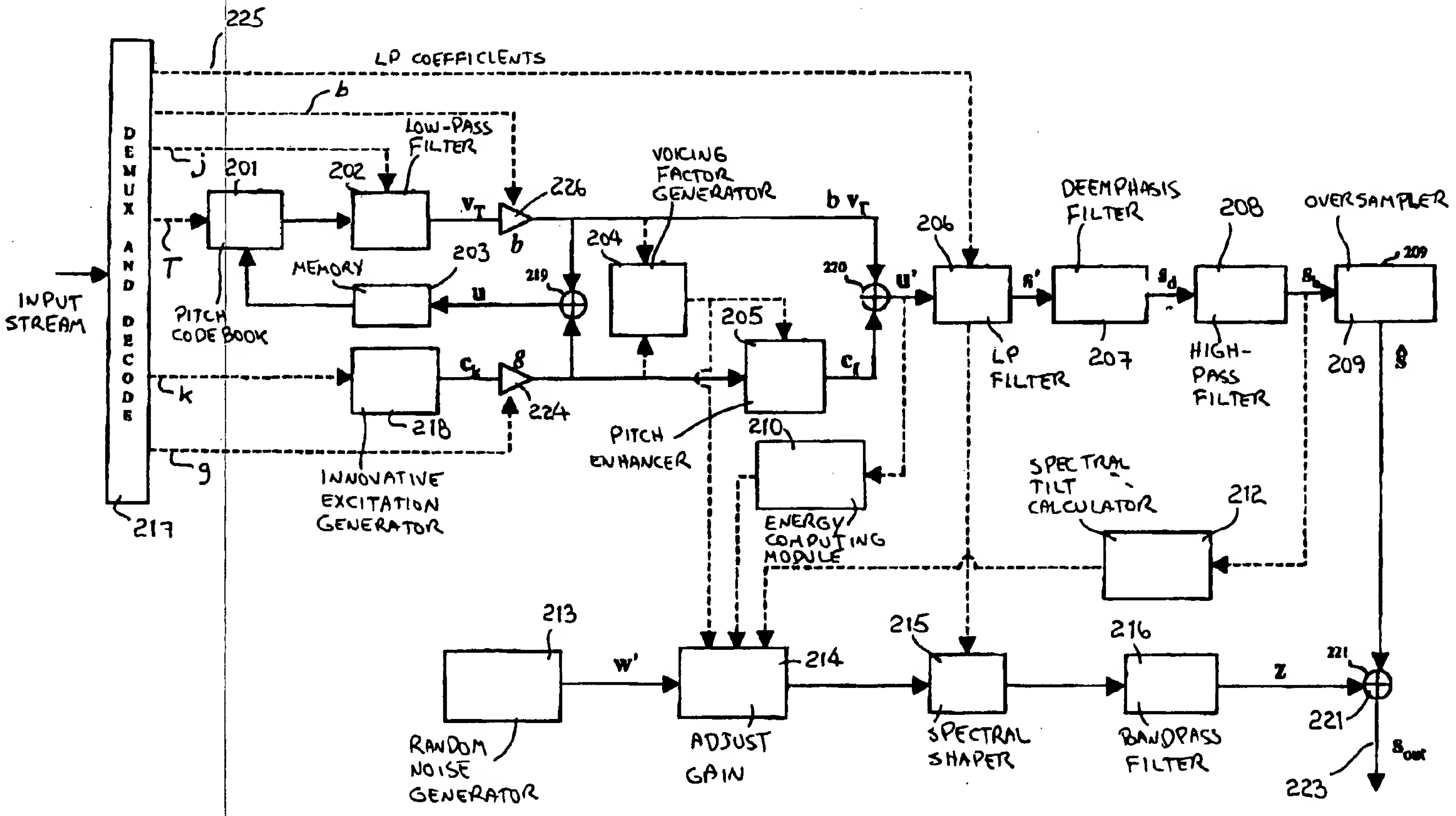
225
LP COEFFICIENTS
b
DEMUX AND DECODE
j
201
202
LOW-PASS FILTER
v_T
226
b
VOICING FACTOR GENERATOR
$b\,v_T$
206
DEEMPHASIS FILTER
208
OVERSAMPLER
209
INPUT STREAM
T
PITCH CODEBOOK
MEMORY
203
u
219
204
220
u'
205
$\hat{s}'$
s_d
s_h
LP FILTER
207
HIGH-PASS FILTER
209
$\hat{s}$
k
c_k
g
224
218
c_f
PITCH ENHANCER
210
g
217
INNOVATIVE EXCITATION GENERATOR
ENERGY COMPUTING MODULE
SPECTRAL TILT CALCULATOR
212
213
w'
214
215
216
z
221
221
RANDOM NOISE GENERATOR
ADJUST GAIN
SPECTRAL SHAPER
BANDPASS FILTER
s_{out}
223
OUTPUT SPEECH

Figure 2

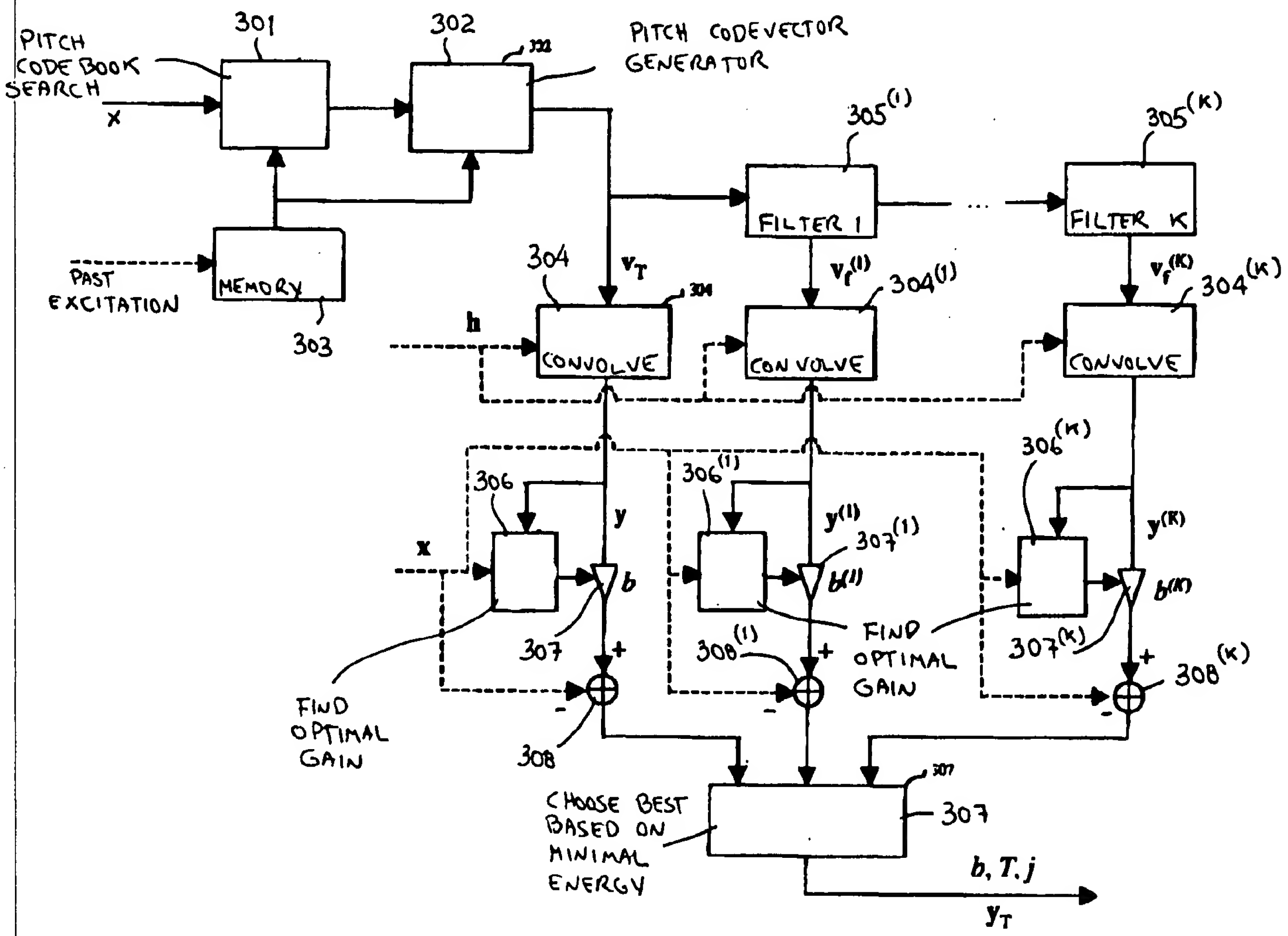
PITCH CODE BOOK SEARCH
x
301
302
302
PITCH CODEVECTOR GENERATOR
PAST EXCITATION
MEMORY
303
305$^{(1)}$
305$^{(K)}$
FILTER 1
FILTER K
...
304
304
v_T
$v_f^{(1)}$
304$^{(1)}$
$v_f^{(K)}$
304$^{(K)}$
h
CONVOLVE
CON VOLVE
CONVOLVE
306
306$^{(1)}$
306$^{(K)}$
x
y
$y^{(1)}$
$y^{(K)}$
b
$b^{(1)}$
$b^{(K)}$
307
307$^{(1)}$
307$^{(K)}$
+
-
308
308$^{(1)}$
308$^{(K)}$
FIND OPTIMAL GAIN
FIND OPTIMAL GAIN
307
307
CHOOSE BEST BASED ON MINIMAL ENERGY
b, T, j
y_T

Figure 3